PACIFIC SOUTHWEST
Forest and Range
Experiment Station

FOREST SERVICE
U.S.DEPARTMENT OF AGRICULTURE
P.O. BOX 245, BERKELEY, CALIFORNIA 94701

TWO-STAGE STRATIFIED SAMPLING TO ESTIMATE HERBAGE YIELD

C. Eugene Conrad William G. O'Regan

USDA Forest Service
Research Note PSW-278
1973

Abstract: A two-stage stratified procedure for sampling herbage yield is described. In stage one, plots are stratified; in stage two, selected plots are clipped, and the clippings are weighed. Equations were developed to calculate the mean and variance of the mean. To achieve minimum variance when time for sampling is limited, equations were developed for determining the sample size and the distribution of the plots between stages one and two. The procedures were demonstrated by using data from the San Joaquin Experimental Range, in central California. The results showed that two-stage sampling could produce a 15 percent reduction in variance of mean weight compared with that from simple random sampling.

Oxford: 268.5–015.5.
Retrieval Terms: range sampling; herbage yield; statistical samples; two-stage sampling; San Joaquin Experimental Range.

The effectiveness of sampling to estimate herbage yield of plant species can often be improved by using a two-stage stratified procedure. In stage one, many plots are stratified according to some easily observed characteristic, such as species composition of foliar cover. In stage two, selected plots are measured in detail by clipping the vegetation, sorting the clippings into species, and weighing the sorted clippings.

This note presents statistical models that describe the population mean and variance of the mean where two-stage stratified sampling is used. Equations are given to use with sample data for calculating the estimate of the population mean and variance of the mean. Equations are also given for determining sample size when the budget limits the time available for sampling. A method for determining the number of plots that should be taken in each of the two stages of sampling is given.

The models and procedures were adapted from Horwitz.[1] Several of the equations were also developed independently by Cochran.[2] How the procedures can be applied is illustrated by using data from a study on the San Joaquin Experimental Range, in central California. Two-stage stratified sampling can be adapted for use in other situations in which information can be subdivided into strata.

PROCEDURE

We used a 45° inclined point quadrat[3] to determine the species composition of the foliar cover in stage one of sampling.[4] A point quadrat consists of a frame that acts as a guide when the operator is lowering pointed pins through the plant foliage to the ground surface. We designed a frame with eight points on each 1-sq.-ft. plot. The points were lowered through a 45° path. Compared with the angle obtained in using vertical pins, this angle increases the chances of hitting the foliage directly—especially where grass is involved.

As a pin was lowered through the foliage we recorded the first species hit. If no direct hit was made, we recorded the species nearest the path of the point. Thus each species was tallied from zero to eight times for each plot. The final step in first-stage sampling consisted of constructing a frequency distribution of nine strata described by the 0 to 8 tallies per plot for each species.

In stage two, some of the plots sampled in stage one were clipped. The clippings were air-dried in a glass house, sorted, and weighed, by species. Compared with the cost of first-stage sampling, clipping the plots and sorting and weighing clippings is expensive. Part of the sampling effort is devoted to estimating the size of the strata in the first-stage and part to measuring herbage yield by strata in the second-stage.

The plots observed in the first- and second-stages of sampling were randomly placed according to a scheme described by Reppert, *et al.*[5]

THE MODEL

The population we were sampling was the herbaceous vegetation growing on "A" elements of a range site on the San Joaquin Experimental Range, in central California. Each element is potentially a 1-square-foot plot. Selection of plots in stage one is assumed to be random with equal probability of selecting an element. The vegetation on each element selected as a plot is described in terms of cover composition; that is, hits per species, and in terms of weight per species (W) as defined below.

The symbols used to describe the population are defined as follows:

A = number of elements in the population

n = number of strata (number of pins + 1)

A_h = number of elements in the h^{th} stratum where $h = 0, 1, \ldots, n$

$\Pi_h = \frac{A_h}{A}$ = proportion of the area in the h^{th} stratum

W_{hi} = weight of the given species on the i^{th} element in the h^{th} stratum

$\theta_h = \frac{1}{A_h} \sum_{i=1}^{A_h} W_{hi}$ = actual mean weight/element in the h^{th} stratum

$\theta = \sum_{h=0}^{n} \Pi_h \theta_h = \sum_{h=0}^{n} \frac{A_h}{A} \sum_{i=1}^{A_h} \frac{W_{hi}}{A_h} = \frac{1}{A} \sum_{h=0}^{n} \sum_{i=1}^{A_h} W_{hi}$

= actual mean weight/element in the population for the given species

$\sigma_h^2 = \frac{1}{A_h} \sum_{i=1}^{A_h} (W_{hi}-\theta_h)^2$ = actual variance of the h^{th} stratum

$\sigma^2 = \sum_{h=0}^{n} \Pi_h \sigma_h^2 + \sum_{h=0}^{n} \Pi_h (\theta_h - \theta)^2$ = actual variance of the population

C_N = cost per plot for first stage sampling

C_M = cost per plot for second stage sampling

The symbols used to describe the sample from the population were defined as follows:

N = number of plots chosen at random from A

N_h = observed number of plots in the h^{th} stratum

M = number of plots chosen from N for second stage sampling

M_h = number of plots chosen at random from N_h for second stage sampling

$P_h = \frac{N_h}{N}$ = proportion of the sampled plots in h^{th} stratum

$\hat{\theta}_h = \frac{1}{M_h} \sum_{i=1}^{M_h} W_{hi}$ = estimated mean weight in h^{th} stratum

$$\hat{\theta}_S = \sum_{h=0}^{n} \frac{N_h}{N} \hat{\theta}_h = \frac{1}{N} \sum_{h=0}^{n} \frac{N_h}{M_h} \sum_{i=1}^{M_h} W_{hi} \qquad (1)$$

= estimated mean weight per plot for the given species.

Horwitz[1] has shown that the variance of $\hat{\theta}_S$ is

$$V(\hat{\theta}_S) = \frac{1}{N} \sum_{h=0}^{n} \frac{\Pi_h \sigma_h^2}{M_h} [1 + (N-1)\Pi_h] + \frac{1}{N} \sum_{h=0}^{n} \Pi_h (\theta_h - \theta)^2 \qquad (2)$$

and that it has an unbiased estimator

$$\hat{V}(\hat{\theta}_S) = \frac{1}{N-1} \left[\sum_{h=0}^{n} \frac{P_h S_h^2}{M_h} (N_h - 1) + \sum_{h=0}^{n} P_h (\hat{\theta}_h - \hat{\theta})^2 \right] \qquad (3)$$

in which

$$S_h^2 = \frac{1}{M_h - 1} \sum_{i=1}^{M_h} (W_{hi} - \hat{\theta}_h)^2$$

We assumed the number of elements drawn from each of the strata was small relative to the total number and, therefore, did not include a correction for finiteness. Equation 2 agrees with Cochran's[2] equation 12.6 and provided the finite population correction is ignored, equation 3 agrees with his equation 12.12.

If the number of plots chosen in the second-stage sample are allocated proportionally,

$$M_h = \frac{N_h}{N} M = P_h M,$$

and equation 1 reduces to equation 4 for the estimated mean weight per plot

$$\hat{\theta}_S = \frac{1}{M} \sum_{h=0}^{n} \sum_{i=1}^{M_h} W_{hi} \qquad (4)$$

The variance of $\hat{\theta}_S$ under the special condition that M is proportionally allocated cannot be directly obtained from equation 2. Horwitz[1] showed that equation 5 is the variance where sampling is proportional

$$V(\hat{\theta}_S) = \frac{1}{M} \sum_{h=0}^{n} \Pi_h \sigma_h^2 + \frac{1}{N} \sum_{h=0}^{n} \Pi_h (\theta_h - \theta)^2 \qquad (5)$$

$$= \frac{V_M}{M} + \frac{V_N}{N} \qquad (6)$$

V_M and V_N in equation 6 are equivalent to the first and second parts of 5 and are expressions for within and between strata variability, respectively. An unbiased estimate of the variance of $\hat{\theta}_S$ under proportional allocation of M is given by

$$\hat{V}(\hat{\theta}_S) = \frac{1}{N-1} \frac{N}{M^2} \sum_{h=0}^{n} (MP_h - 1) S_h^2$$

$$+ \frac{1}{N-1} \sum_{h=0}^{n} P_h (\hat{\theta}_h - \hat{\theta})^2 \qquad (7)$$

The variance given by 5 or 6 can be minimized, subject to a budget restriction (equation 8)

$$B = MC_M + NC_N \qquad (8)$$

A minimum B is obtained by using Lagrange multipliers in this form:

$$\frac{V_M}{M} + \frac{V_N}{N} - \lambda(B - MC_N - NC_N)$$

We take partial derivatives with respect to M and N, set them to 0 and solve for M and N. The result is substituted into the budget equation (8). After some algebraic manipulation, the following expressions for M and N are derived.

$$M = B\sqrt{V_M/C_M} \; / (\sqrt{C_N V_N} + \sqrt{C_M V_M})$$

$$N = B\sqrt{V_N/C_N} \; / (\sqrt{C_N V_N} + \sqrt{C_M V_M})$$

Since the denominators in the results for M and N are the same, the following equations are equivalent.

$$M = \frac{1}{K} B\sqrt{V_M/C_N} \qquad (9)$$

$$N = \frac{1}{K} B\sqrt{V_N/C_N} \qquad (10)$$

in which

$$K = \sqrt{C_N V_N} + \sqrt{C_M V_M}$$

To arrive at reliable values for M and N, acceptable values for the variances (V_M and V_N) and costs (C_M and C_N) are assumed. Normally these values can be obtained from previous sampling. The reliability will, of course, be improved as feed-back information from sampling becomes available.

When expressions 9 and 10 are substituted for M and N in equation 6 we find:

$$V(\hat{\theta}_S) = \frac{1}{B} K(\sqrt{V_M C_M} + \sqrt{V_N C_N}) \qquad (11)$$

$$= \frac{1}{B} K^2$$

SAN JOAQUIN STUDY

At the San Joaquin Experimental Range, we used 1-square-foot plots to sample herbage yield, including the yield of individual annual plant species. A frequency distribution of the number of plots falling into each of nine strata was obtained in stage one. Species weight was determined in stage two.

Table 1–*Simulated population values for* Bromus mollis *obtained from 310 plots measured by using a two-stage stratified sampling procedure at the San Joaquin Experimental Range, California*

Parameter	Strata (h)									
	0	1	2	3	4	5	6	7	8	Total
π_h [1]	0.289	0.212	0.161	0.121	0.088	0.061	0.039	0.021	0.008	1.000
θ_h [2]	1.350	3.375	4.875	8.170	9.400	9.912	12.693	13.385	15.900	5.214
σ_h^2 [3]	3.551	11.886	14.758	33.310	32.199	31.214	54.442	87.682	40.330	32.446

[1] π_h = proportion of the population in the h^{th} stratum.
[2] θ_h = actual mean weight in grams/element in the h^{th} stratum.
[3] σ_h^2 = variance of the h^{th} stratum.

By using the data collected, we computed reasonable values for a simulated population of *Bromus mollis* (soft chess) *(table 1).* The cost to observe and record the point quadrat data (C_N) was 4 man-minutes per plot. The cost to clip, sort, and weigh the herbage for species weight composition (C_M) was 68 man-minutes per plot. We used the proportions, means, variances, and costs as though they were the actual values of the parameters describing a large population. Since the numbers were computed from a sample drawn from a population, we have labeled the results "simulated population values." The purpose of the simulated values was to enable us to determine the efficiency of two-stage stratified sampling by comparing the variances resulting from the use of two-stage sampling and simple random sampling.

Calculations using data in *table 1* showed these results:

$$V_N = \sum_{h=0}^{n} \Pi_h \sigma_h^2 = 13.489$$

$$V_M = \sum_{h=0}^{n} \Pi_h (\theta_h - \theta)^2 = 18.956$$

$$K = \sqrt{V_M C_M} + \sqrt{V_N C_N} = 43.249$$

If we had fixed the budget (B) at 6,000 man-minutes (100 man-hours), the minimum variance calculated using equation 11 would have been 0.312. This is the minimum variance for the simulated population given in *table 1.* Computation using equation 10 and our simulated population values showed that 255 plots (N) should be drawn from "A" elements for first-stage sampling, and computation using equation 9 showed that 73 plots (M) should be drawn from N for second-stage sampling. If the values for M and N are included in equation 8, 16 man-minutes of the allotted 6,000 would be unused because of rounding the results of equations 9 and 10. We missed achieving the minimum variance as a result. If the 16 minutes were arbitrarily used to increase first-stage sampling by four plots, the calculated variance would be close to the minimum. For comparison a simple random sample of size M′ would be B/C_M and has a variance

$$V(\hat{\theta}_R) = \frac{\sigma^2}{M'} = \frac{(V_N + V_M)C_M}{B}$$

$$= \frac{32.445(68)}{6000} = 0.368 \qquad (12)$$

Comparing the results showed that the minimum variance using two-stage stratified sampling was 15 percent less than that from simple random sampling.

With a fixed budget (B) the difference between the variance for simple random sampling (equation 12) and two-stage sampling (equation 11) is

$$V(\hat{\theta}_R) - V(\hat{\theta}_S) = \frac{1}{B} V_N (C_M - C_N) - \frac{2}{B} \sqrt{V_M C_M V_N C_N}$$

This algebraic result leads to these observations. First, the numerical size of the difference decreases as B increases; in other words, if a range scientist is operating with a small budget, two-stage sampling may have more value than if his budget is large. Second, the cost of a first-stage plot (C_N) will be less than the cost of a second-stage plot (C_M) if $V(\hat{\theta}_S)$ is less than $V(\hat{\theta}_R)$.

The difference between $V(\hat{\theta}_R)$ and $V(\hat{\theta}_S)$ decreases as V_N becomes smaller and V_M becomes larger. V_M is essentially a within-stratum variance, and V_N is an among strata variance. If the strata means are equal, V_M is the same as the over-all variance. If a first-stage sample is poorly designed or not

related to the values measured in the second-stage, V_N approaches 0 and the variance is accounted for by V_M. But regardless of the effectiveness of the two-stage procedure, the mean and variance have unbiased estimators.

A problem alluded to in the discussion of proportional allocation was the error resulting from rounding M and N to the nearest whole number. In practice, rounding M and N appears to cause only an insignificant effect on $V(\hat{\theta}_S)$. A similar problem arises when one attempts proportional distribution of the M plots among the strata. Rounding will probably be called for in each stratum, consequently the arithmetical difficulty introduced by rounding is compounded.

We used proportional sampling to simplify the job of minimizing the variance subject to a restricted budget. But as pointed out above, P_hM will rarely be a whole number. Also P_hM may be a value less than 2 which is the minimum needed for determining S_h^2. Use of the generalized equations, 2 and 3 avoids the arithmetic problem created by rounding under proportional allocation. By using data from *table 1*, solution of equation 2, in which M_O is set at 20 and M_8 is set at 2, gives $V(\hat{\theta}_S) = 0.320$. The indicated increase in precision over $V(\hat{\theta}_R)$ is 13 percent. No difficulty is introduced—provided selection of the M_h plots is random.

NOTES

[1]Horwitz, Leah. 1959. (Unpublished notes on file at Pacific Southwest Forest and Range Experiment Station, Berkeley, Calif.)

[2]Cochran, William G. *Sampling techniques.* Ed. 2. New York: John Wiley and Sons, Inc. p. 328. 1963.

[3]Drew, W. B. *Studies on the use of the point-quadrat method of botanical analysis of mixed pasture vegetation.* J. Agr. Res. 69:289-297. 1944.

Wilson, J. Warren. *Inclined point quadrats.* New Phytol. 59: 1-8. 1960.

[4]Reppert, Jack N., Meredith J. Morris, and Charles A. Graham. *Estimation of herbage on California annual-type range.* J. Range Manage. 15:319-323. 1962.

[5]Reppert, Jack N., Merton J. Reed, and Pinhas Zuzman. *An allocation plan for range unit sampling.* J. Range Manage. 15: 190-193. 1962.

The Authors

C. EUGENE CONRAD is a research botanist assigned to the Station's work unit on flood and sediment reduction in steep unstable brushlands, with headquarters at Glendora, California. He is a graduate of Oregon State College (B.S. degree in general agriculture, 1956; M.S. degree in range management, 1959). He joined the Forest Service, and the Station staff in 1961. **WILLIAM G. O'REGAN**, a mathematical statistician, is chief of the Station's biometrics branch, with headquarters in Berkeley. He earned B.S. (in agricultural economics, 1949) and Ph.D. (1962) degrees at the University of California, Berkeley. He has been a member of the Station staff since 1957.

The Forest Service of the U.S. Department of Agriculture

. . . Conducts forest and range research at more than 75 locations from Puerto Rico to Alaska and Hawaii.

. . . Participates with all State forestry agencies in cooperative programs to protect and improve the Nation's 395 million acres of State, local, and private forest lands.

. . . Manages and protects the 187-million-acre National Forest System for sustained yield of its many products and services.

The Pacific Southwest Forest and Range Experiment Station

represents the research branch of the Forest Service in California and Hawaii.

GPO 794–424/3729